ACADEMIC PRESS

Available at WWW.MATHEMATICSWEB.ORG POWERED BY SCIENCE DIRECT®

Journal of Multivariate Analysis 84 (2003) 387–402

Journal of Multivariate Analysis

http://www.elsevier.com/locate/jmva

Bayesian networks for discrete multivariate data: an algebraic approach to inference

J.Q. Smith and J. Croft*,[1]

Department of Statistics, University of Warwick, Coventry CV4 7AL, UK

Received 2 January 2001

Abstract

In this paper we demonstrate how Gröbner bases and other algebraic techniques can be used to explore the geometry of the probability space of Bayesian networks with hidden variables. These techniques employ a parametrisation of Bayesian network by moments rather than conditional probabilities. We show that whilst Gröbner bases help to explain the local geometry of these spaces a complimentary analysis, modelling the positivity of probabilities, enhances and completes the geometrical picture. We report some recent geometrical results in this area and discuss a possible general methodology for the analyses of such problems.

AMS 1991 subject classifications: 62H99; 62-09; 62F15; 13P10

Keywords: Graphical models; Bayesian networks; Hidden variables; Gröbner basis; Latent class analysis

1. Introduction

Recently, Pistone et al. [26] reviewed how techniques from algebraic geometry—especially those associated with Gröbner bases—could be usefully employed in the study of certain classes of statistical models associated with design (for example, [14,25]), reliability theory (for example [10,13,24]) and statistical distribution theory (for example [8,9,27]). In this paper we will discuss how these algebraic techniques can also be helpful in elucidating, manipulating and estimating probabilities in

*Corresponding author.

E-mail addresses: j.q.smith@warwick.ac.uk (J.Q. Smith), jcc@ukonline.co.uk, jcc@stats.warwick.ac.uk (J. Croft).

[1] Thanks to the EPSRC for funding under GR/M56005 101.

0047-259X/03/$ - see front matter © 2003 Elsevier Science (USA). All rights reserved.
doi:10.1016/S0047-259X(02)00067-2

discrete Bayesian networks when some of the variables in the network are never observed.

It is straightforward to estimate probabilities on a Bayesian network when all the variables in the model are observed. These models form a curved exponential family [18]. A simple Bayesian analysis is available using a product Dirichlet conjugate family on the conditional probabilities defining the model (for example, [3,19,34]). This allows a simple closed form prior to posterior analysis to take place. There are also various log-linear parametrisations of these models (for example, [21,35,39]) that allow for more flexible accommodation of model and prior information as this is appropriate.

However, when data on some of the variables is missing the probabilities in even relatively simple Bayesian networks can become very difficult to reliably estimate. Only a very small subset of these models form exponential families [12,32]. The joint probabilities in the probability tables of the observed variables therefore do not typically lie on a smooth manifold and so the corresponding likelihood can be geometrically very strange. In fact, we will demonstrate below that even in simple models such a likelihood can have multiple disconnected global maxima.

There are various unpleasant consequences of this feature. First, conventional model selection, such as BIC [30] or chi-squared/divergence methods [39], are no longer demonstrably valid model selection techniques, even in the limit. Second, many of the models become unidentifiable. This means that rather innocuous looking features of a prior distribution, like its tail area characteristics, chosen for convenience rather than faithfulness, can have a strong effect on the resulting posterior distribution. Furthermore, because of the complicated form of the likelihood these strong effects tend to be unpredictable. Finally, because the likelihood and posterior density typically have many disconnected global maxima standard numerical techniques such as MCMC and EM algorithms [1,3] tend to breakdown—appearing to converge to a bogus posterior sample density or picking just one of many local maxima of the likelihood.

Recognising these difficulties Cowell et al. [4] proposed a methodology based on choosing a prior so that the predictive distribution is approximated well. For example Cowell [2], the variational approach seen in [23] and the interval probability approach in [28]. Such methods are appropriate if the Bayesian networks are used for prediction of the observable variables. However, it is common for the model parameters themselves to be the intrinsic quantities of interest—like the probability a certain disease will cause a particular symptom, as seen in the paper by Spiegelhalter and Cowell [33]. In such cases there is no simple way of avoiding the difficult geometry of the model parameters by using these types of approximation methods. If the models are used to estimate probabilities associated with hidden variables in the system—as is the case in [33]—at best these methods essentially deny any possibility of learning about probabilities with geometrically non-trivial likelihoods from the data, as was the conclusion of these authors.

As far as we are aware, all the current studies of the geometry of the parameter spaces of Bayesian networks with hidden variables assume that the cell probabilities of the observed variables are known. Furthermore, most of these papers study

certain simple Bayesian networks with hidden variables (for example [12,3]) and then only to demonstrate why standard inferential techniques can no longer be expected to work for this class of models. The class of latent class models, which can be thought of as a very simple Bayesian network with one hidden variable, is the only exception to this gap in algebraic estimation theory (see [7,38]). Even here the emphasis is to find a single good estimate of the probability tables or to provide a numerically generated estimate of the posterior distribution of these quantities [11]. Very little attention has been given to understanding the geometry underpinning this estimation problem, classifying the types of ambiguities inherent in any good inferential procedure and studying the nature of the instability of the numerical techniques employed.

In this paper we suggest how general techniques from algebraic geometry, and particularly Gröbner bases can be used in a systematic fashion to explore the identifiable regions of the probabilities determining a given discrete Bayesian network. In Section 5 we then make some tentative steps towards developing this into a formal framework for estimating such models.

It is helpful at this stage to introduce an example which illustrates how these ambiguities can occur and why they are intrinsically linked to the geometry of high dimensional polynomials. Thus consider the Bayesian network whose DAG (directed acyclic graph) is given in Fig. 1. The dark nodes S_1, S_2, S_3, S_4 are all observed with S_1, S_2, S_3 being binary with states $\{-1, 1\}$ and S_4 taking four possible values. The light nodes C_1, C_2 are those random variables which are hidden, C_1 begin binary on $\{-1, 1\}$ and C_2 taking 3 possible values. As is well known, this DAG tells us that the joint mass function of these six variables factorises according to the formula

$$p(s_1, c_1, s_2, s_3, c_2, s_4) = p_1(s_1)p_2(c_1|s_1)p_3(s_2|c_1)p_4(s_3|c_1)p_5(c_2|s_1, s_2)p_6(s_4|c_2),$$

where each function $p(x \mid y)$ is a function of its arguments x, y only and where for each value y_j of the conditioning variable(s)

$$\sum_{x_i} p(x_i|y_j) = 1$$

and for each argument (x_i, y_j),

$$p(x_i|y_j) \geqslant 0.$$

In problems like this we have two tasks. The first is to estimate, as far as we can, the vector of all these probabilities from the data set provided and so quantify the

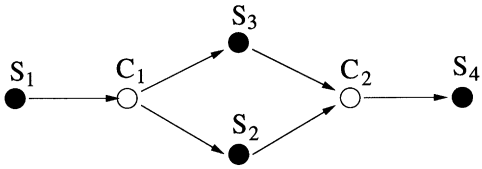

Fig. 1. An example of a Bayesian network with observed and hidden variables.

causal strengths of different components of the model given that the posited model is valid. The second is to understand when the given data set might suggest that the conditional independence structure posited in the model is suspect. We will see that both these problems relate to the geometry of the mass function $p(.)$ above. Some aspects of this implied geometry can be usefully analysed using Gröbner bases whilst others need to address extra convexity issues implicit in this formulation. The added complexity of the geometry is solely due to the existence of the inequality constraints above—particularly those involving the hidden variables in the problem.

If all the variables are observed it is easy to see that the likelihood is monomial. It is precisely for this reason that the model is accessible to a conjugate prior to posterior analysis or to a log-linear analysis.

However when the Bayesian networks contain hidden variables estimation is more difficult. The likelihood will then be a function of the joint marginal distribution of the manifest variables. In our example the density from which the likelihood derives is given by

$$p_0(s_1, s_2, s_3, s_4) = \sum_{c_1, c_2} p(s_1, c_1, s_2, s_3, c_2, s_4).$$

Such a density will typically also factor, but usually more coarsely—see [32].

An appropriate factorisation of the problem above is given by

$$p_0(s_1, s_2, s_3, s_4) = p_1(s_1) p_7(s_2, s_3|s_1) p_8(s_4|s_2, s_3),$$

where

$$p_7(s_2, s_3|s_1) = \sum_{c_1} p_2(c_1|s_1) p_3(s_2|c_1) p_4(s_3|c_1)$$

and

$$p_8(s_4|s_2, s_3) = \sum_{c_2} p_5(c_2|s_2, s_3) p_6(s_4|c_2).$$

What makes these sorts of models more challenging is the fact that some of the components of the factorisation—in this case $p_7(.)$ and $p_8(.)$—are no longer monomial in the conditional probabilities of interest. There is now no obvious transformation, such as taking logarithms, to simplify these functions. The intrinsic geometry of these functions can therefore be very rich.

Because we often still have a factorisation of the likelihood/joint density over the manifest variables, to understand the geometry and estimate the full DAG model it is sufficient to focus on its factors. Thus to understand $p_0(.)$ we need only come to a geometrical understanding of its factors $p_1(s_1), p_7(s_2, s_3|s_1), p_8(s_4|s_2, s_3)$. These factors are themselves DAG models, see for example Fig. 2. Unfortunately, some very simple factors are associated with non-trivial geometry.

Geometrically interesting problems can arise only for collections of conditional probabilities having a missing variable as one of their arguments. In our example these are $\{p_2(c_1|s_1), p_3(s_2|c_1), p_4(s_3|c_1), p_5(c_2|s_2, s_3), p_6(s_4|c_2)\}$. These parameters will typically be unidentifiable in particularly unpleasant ways which we will illustrate below.

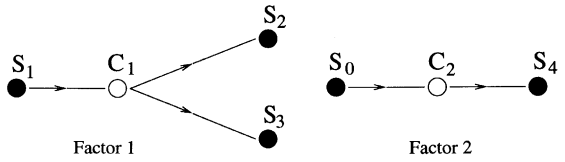

Fig. 2. A decomposition of a Bayesian network into simple factors where $S_0 = S_2 + 2S_3$.

In the next section we shall review some of the basic theory of Gröbner bases we need in the paper. Gröbner bases are much easier to interpret and derive from computer algorithm packages if a parametrisation is chosen in a way sympathetic to any of the inherent symmetries of the model. In Section 3 we review a parametrisation in terms of moments that we have found helpful for addressing these symmetries and hence allowing the current software to be used in moderate sized problems of this kind.

We demonstrated above how a larger graphical model can be decomposed into more simple factors each with their own geometry. In Sections 4 and 5 we will examine these two factors whose geometry is now relatively well understood. We use these components to illustrate the general point of this paper that the algebra associated with Gröbner bases needs to be used in conjunction with other geometrical methods to allow the analysis of the convexity constraints arising from the demand that probabilities must be positive. We argue that together the algebraic and convex geometry combine to give a complete picture of the geometry of the likelihoods arising from Bayesian networks with hidden variables.

In Section 5 we report on the statistical implications of some new results about the geometry of stochastically factorisable spaces derived by Mond et al. [22]. This paper gives the proof of a classification of the possible geometry arising from the convexity constraints of the second component of our illustrative Bayesian network in terms of its possible homotopy types. In Section 5 and the appendix we also describe a new and fast algorithm—using Gröbner bases—to estimate various parameters in this model.

We begin the paper with a brief introduction to Gröbner bases.

2. Gröbner bases and systems of simultaneous polynomial equations

It is common in statistics to be faced with the task of simplifying or even solving sets of simultaneous polynomial equations in many parameters/variables. A natural first approach to this task is to rearrange, eliminate and then substitute as we would in Gaussian elimination for a linear system. Thus, we might attempt to simplify the system by manipulation until one equation is immediately soluble, then deduce the rest of the solutions by substitutions.

However, in polynomial systems, it is apparent that if we do this unsystematically, even for problems with only a few variables and equations, this naive approach contains too many choices or permutations of manipulation to be quick or even feasible.

To develop a systematic approach it is useful to employ more general theory on the underlying algebraic structures. What follows is a very brief summary of the approach to these systems of equations and some references to a more complete exposition.

Let $\mathbb{C}[x_1, \ldots, x_n] = \mathbb{C}(\underline{x})$ be the set of all polynomials in the variables $x_1, \ldots, x_n$ with coefficients in $\mathbb{C}$. An *ideal* I is a subset of $\mathbb{C}(\underline{x})$ closed under summation and products of elements in $\mathbb{C}(\underline{x})$. That is for all $f, g \in I$ and $h \in \mathbb{C}(\underline{x})$ we require that

$$f + g \in I \quad \text{and} \quad hf \in I.$$

An ideal I is *finitely generated* if there exists $f_1, \ldots, f_s \in \mathbb{C}(\underline{x})$ which satisfies: $f \in I$ implies that there exists $h_1, \ldots, h_s$ such that

$$f = \sum_{i=1}^{s} h_i f_i.$$

We write $I = \langle f_1, \ldots, f_s \rangle$ and $\{f_1, \ldots, f_s\}$ is called a basis for I. Such bases are not unique. A Gröbner basis $\{g_1, \ldots, g_t\}$ for an ideal I is defined to be a basis such that $\langle lm(g_1), \ldots, lm(g_t) \rangle = \langle lm(f) : f \in I \rangle$ where $lm(f)$ denotes the *leading monomial* of a polynomial f with respect to some *monomial ordering*. A monomial ordering, $\prec$, is simply an ordering on the set of monomials or product terms. For example, in this article we use the *lexicographical* ordering on $\mathbb{C}[x_1, \ldots, x_n]$ defined by

$$x_1^{i_1} \cdots x_n^{i_n} \prec x_1^{j_1} \cdots x_n^{j_n}$$

iff $i_1 = j_1, \ldots, i_k = j_k, i_{k+1} < j_{k+1}$ for some k. In other words, the first variable in the list $x_1, \ldots, x_n$ with different exponents has lower exponent in the *lesser* monomial (with respect to $\prec$).

Let $\sigma(\langle f_1, \ldots, f_s \rangle)$ be the set of points $c = (c_1, \ldots, c_n) \in \mathbb{C}^n$ for which $f \in \langle f_1, \ldots, f_s \rangle \Rightarrow f(c) = 0$.

Now let $f_1 = \cdots = f_m = 0$ be a system of polynomial equations in the variables $x_1, \ldots, x_n$. It is easy to show that the set of solutions to the system is exactly $\sigma(\langle f_1, \ldots, f_m \rangle)$.

The idea behind finding this solution set is to decompose the ideal $\sigma(\langle f_1, \ldots, f_m \rangle)$ into intersections of *primary* ideals. These primary ideals correspond in some sense to single solutions for certain variables. These solutions can then be substituted to uncover the corresponding solutions for other variables. This idea lies behind the definition of a Gröbner basis and indeed is used to form an algorithm for finding them.

After finding a Gröbner basis $\{g_1, \ldots, g_t\}$ such that $\langle f_1, \ldots, f_m \rangle = \langle g_1, \ldots, g_t \rangle$ it is reduced with special reference to the monomial term ordering. This reduced basis now contains a polynomial in just one variable which can be solved by standard techniques. With this solution, another element of the reduced basis can be used after substitutions to find corresponding solutions for other variables. This continues until all solutions are found.

Because this method is automatic it can be found in many computer algebra packages. The most useful of these we found to be *Maple* and *CoCoa*.

Example. Consider the system of equations

$$x_1^2 + x_2^2 + x_3^2 = 3,$$

$$x_1 x_2 x_3 - 1 = 0,$$

$$x_1 x_3 + x_2 - 2 = 0.$$

The ideal

$$\langle x_1^2 + x_2^2 + x_3^2 - 3, x_1 x_2 x_3 - 1, x_1 x_3 + x_2 - 2 \rangle$$

has, with lexicographical monomial term ordering, the reduced Gröbner basis

$$\{x_1 + x_2 x_3 - \tfrac{1}{2x_3^3} + 2x_3^3 - \tfrac{7}{2x_3}, x_2^2 - 2x_2 + 1,$$
$$x_2 x_3^3 - x_2 + \tfrac{1}{2x_3^4} - 2x_3^2 + \tfrac{3}{2}, x_3^6 - 3x_3^4 + 3x_3^2 - 1\}.$$

From the last basis element we get the solutions $x_3 = \pm 1$. Substituting $x_3 = 1$ leads the other equations to the solution $x_1 = x_2 = 1$ and substituting $x_3 = -1$ leads to the solution $x_1 = -1, x_2 = 1$. Hence the system has two solutions $(1,1,1)$ and $(-1,1,-1)$. Gröbner basis techniques can also be used to show the multiplicity of both roots is 4.

For further details on these methods and their wider applications see [2,20,26].

3. Moments instead of probabilities for symmetrical geometry

There are several reasons why parametrising a Bayesian network in terms of its conditional probabilities is geometrically inconvenient. The first and most obvious technical niggle is the condition that probabilities need to sum to one, which implicitly reduces dimensions of spaces by one.

A second more substantial problem is that, in most cases, the constraints imposed on the probabilities by conditional independence statements are too complex to be analysed or simplified even using Gröbner basis methods. An exception is the graphical structure studied in Section 5.

Thirdly, a change in the dimension of a hidden variable within the structure can have a profound effect on the geometrical description of the problem. Within the usual parametrisation it is not always clear why we should expect this sensitivity.

It has been known for some time—see for example [25] or [14]—that there is a linear invertible transformation between the joint probability mass function of a collection of variables on a given sample space and a collection of non-central moments defined by an ideal. Using this simple linear transformation we not only lose the probability summation constraint but also preserve symmetrical relationships on the sample space. In particular, any joint distribution on a subset of the variables will only depend on the joint moments of those variables. Any conditional independence statements can be expressed as sets of quadratic constraints on central moments and so impose lower order polynomial constraints on the space of

moments defined by the sample space of the problem. This is therefore the most natural geometrical parametrisation.

Of course, there is a cost to transforming to a moment representation in this way. The inequality conditions constraining our space of interest (that all probabilities are non-negative) are transformed to rather more complicated conditions. However, if we work with non-central moments, these just define a convex region with linear boundaries. This remains true if, alternatively, we parametrise with central moments by conditioning on the first moments. The price of these additional constraints is therefore usually worth paying.

To illustrate the moment transformation suppose $X_1, X_2, \ldots, X_n$ are all binary random variables whose sample space is $\{-1, 1\}$. Then it is easy to check that

$$\prod_{i=1}^{n} X_i^{b_i} = \prod_{i=1}^{n} X_i^{a_i}$$

for any vector of integers $(b_1, b_2, \ldots, b_n)$ such that $a_i = b_i \bmod 2$. In particular, this tells us that the joint moments—which clearly define this finite discrete joint probability—must satisfy

$$E\left(\prod_{i=1}^{n} X_i^{b_i}\right) = E\left(\prod_{i=1}^{n} X_i^{a_i}\right)$$

Therefore, the $2^n - 1$ non-central moments, where $(a_1, a_2, \ldots, a_n)$ is a non-zero binary string, define the space. For example when $n = 3$, these would be

$$\{E(X_1), E(X_2), E(X_3), E(X_1X_2), E(X_1X_3), E(X_2X_3), E(X_1X_2X_3)\}$$

and these replace the 8 probabilities with their one summation constraint. These non-central moments are linear in the joint probabilities, for example

$$E(X_1) = \sum_{x_2,x_3} (-1)^{(x_1-1)/2} p(x_1, x_2, x_3).$$

In general, discrete joint distributions can be spanned by particular collections of moments which depend in form on the joint sample space of the mass function. In more complicated situations, the necessary moments can be calculated using the Gröbner base technology outlined in Section 2, as was alluded to, but not described, in [31].

It is also straightforward to express conditional independence statements as extra polynomial constraints on moments. In particular, whenever a conditioning variable W is binary, conditional independence gives very simple quadratic relationships in central moments. For example, Settimi and Smith [31] prove that, if $Y_1, Y_2, \ldots, Y_n$, $Z_1, Z_2, \ldots, Z_m, W$ are all binary $\{-1, 1\}$ variables then $(Y_1, Y_2, \ldots, Y_n)$ is independent of $(Z_1, Z_2, \ldots, Z_m)$ given W iff

$$Var(W)\, Cov\left(\prod_{i=1}^{n} Y_i^{a_i}, \prod_{j=1}^{m} Z_j^{b_j}\right) = Cov\left(W, \prod_{i=1}^{n} Y_i^{a_i}\right) Cov\left(W, \prod_{j=1}^{m} Z_j^{b_j}\right)$$

for all $a_i, b_j = 0, 1, 1 \leqslant i \leqslant n, 1 \leqslant j \leqslant m$.

In passing it is important to note that any collection of conditional independence statements—not only those associated with DAGs but also more general structures like MAGs (mixed ancestral graphs, see [29]) which build on collections of conditional independence statements—will be expressible in such families of algebraic equations. So the geometrical techniques discussed here are very generally applicable. Finally, note that by increasing the size of the sample space of the hidden variable we automatically need to introduce polynomial relationships whose terms involve the higher order moments spanning the new space. It is therefore transparent, unlike in the graphical representation of the factorisation of a joint density, that the dimension of the sample space of a hidden variable may have a radical effect on the underlying geometry.

4. Triadic geometry: an example of moment equations derived from sample space and independence assumptions

One conditional independence model which has been studied since Goodman (in [15,16]) is the triadic model with binary variables. One example of this model is the first component given in Fig. 2.

To see how the Gröbner bases work on this example we examine how the mass function $p(s_1, s_2, s_3)$ or equivalently its defining moments are constrained by the algebraic form of the model. Here without loss of generality, and to simplify the analysis below, we will now make the additional assumption that $E(S_1) = E(S_2) = E(S_3) = 0$ and write e.g. $\mu_{12} = E(S_1.S_2)$. It can then be shown that the admissible region for the defining unknown moments $(\mu_{12}, \mu_{13}, \mu_{23}, \mu_{123})$ of the distribution of the manifest variables is a union of four disconnected regions of strictly positive measure. It is therefore inappropriate to validate the model using routine dimension counting methods such as BIC or chi-squared criteria. For example the appropriate chi-squared test statistic would have to have zero dimension! After simplification using Gröbner basis techniques (performed in the computer package *Maple*) it can be shown that this four-dimensional region can be parametrised by the following four polynomial equations in four unknowns:

$$\mu_{12} = a_1 a_2 a_4^2, \quad \mu_{13} = a_1 a_3 a_4^2, \quad \mu_{23} = a_2 a_3 a_4^2$$

and

$$a_4^2 \mu_{123}^2 - 4(1 - a_4^2)\mu_{12}\mu_{13}\mu_{23} = 0,$$

where a_i are all variationally independent of each other and $-1 \leqslant a_i \leqslant 1$, $i = 1, 2, 3, 4$.

Obviously, these types of equations lead to interesting features even when we ignore the constraint that all the parameters have to be no larger than one in modulus. For example, multiplying the first set of three equations together we see a familiar quadrant condition

$$\mu_{12}\mu_{13}\mu_{23} \geqslant 0.$$

Such types of condition have been known to be present for a long time, see [5,36]. However, some of the most interesting geometry derives from the boundedness of the parameters a_i, $1 \leqslant i \leqslant 4$. The point to make here is that this is not recognised by the algebraic packages and other automatic algorithms. Therefore, these algebraic methods need to be supplemented by geometric methods which incorporate positivity conditions.

The effects of these constraints are comparatively straightforward when the hidden variable has just two states. In this case direct substitution methods can be used as in [31]. It is found that the solution space is made up of 4 disconnected cuspoid regions in which $(\mu_{12}, \mu_{13}, \mu_{23})$ must lie and that the hidden variable is identifiable up to aliasing.

However even if we increase the number of states the hidden variable can take by one the non-algebraic conditions have increasing structure and need more geometrical techniques to analyse their effect. They cannot be explained away by aliasing as suggested by Stevens [37].

5. Projective spaces and sandwiched triangles

In some simple conditional independence structures, for example the second factor from Fig. 2, the number of hidden states can help us to identify the form the conditional independence constraints on $p(s_2, s_3, s_4)$ will take. Because of the unusual algebraic simplicity of this case we are able to work with the more familiar conditional independence probabilities and there is no need to reparametrise with moments. As seen before, the conditional independence structure of the second factor in Fig. 2 requires that

$$p(s_4|s_2, s_3) = \sum_{c_2} p(c_2|s_2, s_3)p(s_4|c_2)$$

for all values of S_2, S_3, S_4 and C_2. Again we are primarily interested in the estimation of the factors $p(c_2|s_2, s_3), p(s_4|c_2)$. The conditional independence constraint can be split into algebraic (polynomial) equality constraints, for which Gröbner basis methods will be useful in estimation, and geometric (linear) inequality constraints. It is simplest to analyse these constraints in matrix notation since we have only two manifest variables and hence the factoring above can be written in matrix form as

$$[S_4|S_2, S_3] = [C_2|S_2, S_3][S_4|C_1],$$

where all the matrices are stochastic, that is all elements are between 0 and 1 and all row sums are 1. Without loss of generality, write the linear invertible transformation $S_0 = S_2 + 2S_3$. In Mond et al. [22] a mathematical characterisation of the geometry of this type of factorisation is analysed in detail. Two conditions must be met in order that the matrix $[S_4|S_0]$ admits such a factorisation when the number of hidden states is 3. Therefore, it is sensible to plan an estimation procedure that deals with each condition in turn.

The first is a rank condition imposed by the dimensions of the factoring matrices. The rank of $[S_4|S_0]$ must therefore be 3. Algebraically, this condition can be expressed as a series of polynomial equality constraints on the probabilities. The polynomials are simply the determinants of the 4×4 submatrices of $[S_4|S_0]$. This uses the basic fact from linear algebra that a $n \times m$ matrix has rank k iff all $(k+1) \times (k+1)$ submatrices have zero determinant.

The second condition summarises all the positivity constraints imposed by working with probabilities. This turns out to be a convexity constraint: it simply demands that the convex hull formed from the points representing the rows of $[S_4|C_2]$ contains the representations of the rows of $[S_4|S_0]$. The convexity condition is most simply expressed in terms of a 2D projective plot. Note that only a 2D representation is necessary since $[S_4|S_0]$ must be stochastic and hence its rows already lie in an affine subspace, thus reducing the dimension by 1 (see Fig. 3).

In the 2D plot shown in Fig. 4 the positivity constraints on the entries of our factoring matrices are satisfied since the triangle formed as the convex hull of the rows of $[S_4|C_2]$ is sandwiched between the outer quadrilateral (the boundary of the feasible region) and the inner quadrilateral (the convex hull of the rows of $[S_4|S_0]$). Each particular sandwiched triangle with labelled vertices corresponds to a distinct, equally likely solution $[C_2|S_0], [S_4|C_2]$ satisfying

$$[S_4|S_0] = [C_2|S_0][S_4|C_2].$$

The vertices of such a sandwiched triangle correspond to the rows or conditional distributions of the factoring matrix $[S_4|C_2]$. Obviously, there are typically a continuum of these solutions hence many more than could be obtained by aliasing—simply relabelling the states of the hidden variable.

When estimating the conditional probability parameters it is helpful to first fit a rank 3 model and then adjust the estimates if necessary so that the second convexity constraint is satisfied. Fitting a rank 3 probability matrix to our observed $[S_4|S_0]$ is equivalent to fitting a rank or correspondence model as described in [17]. Routines are given in [17] to find the maximum likelihood estimate of such models but can be numerically unstable and involve a parametrisation which is redundant for our purposes.

For example, to obtain a maximum likelihood estimate Π of $[S_4|S_0]$ of rank 3 directly we need to perform a constrained minimisation of the likelihood ratio to the unconstrained maximum likelihood estimate of $[S_4|S_0]$. As mentioned above the constraints can be formulated as a system of polynomial equations.

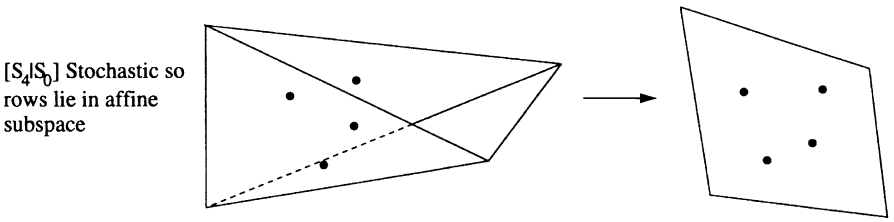

Fig. 3. The origin of the 2D slice plot.

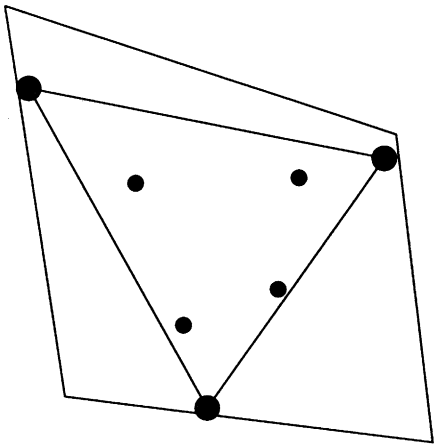

Fig. 4. An example of a sandwiched triangle.

In our example the Lagrangian is given by

$$f(\Pi, \delta, \lambda) = lr([S_4|S_0], \Pi) - \delta\left(\sum_j a_{ij} - 1\right) - \lambda\, det(\Pi).$$

The minimisation is intractable but an iterative algorithm can be derived (see the appendix) which converges quickly to the maximum likelihood estimate of the rank model. At each step in the iteration the solutions to a system of polynomial equations are required. The coefficients of these polynomials depend on the current parameter estimates.

Note that the likelihood ratio of this reduced rank model and the full model on the margins is easily calculated and used to decide whether or not the reduced rank model is a reasonable description of the data.

For the hidden variable model to hold we also need that the additional convexity constraints are satisfied. So having obtained an estimate Π of the desired rank we can now examine the associated 2D (slice) plot to visually test the inequality constraints.

Mond, Smith and van Straten [22] studied this situation geometrically and found that T_v—the space of sandwiched triangles—is (homotopy) equivalent to either a circle or k points where $0 \leqslant k \leqslant 8$. The implication of this is that the constrained likelihood of the conditional probabilities can take very unpleasant forms. Let us examine the possible cases in turn.

When T_v is equivalent to a circle the space of unidentifiability is connected. However the regions of admissable triangles will typically be irregular over the parameter space. So, for example, if a Bayesian uses a routine flat prior distribution over all rank 3 models this will tend to arbitrarily favour some models over others. Interestingly, configurations of points in the 2D representation of the convexity

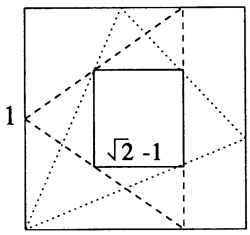

Fig. 5. A degenerate case with 8 separated explanations.

constraints giving rise to this case often suggest that a simpler statistical model, typically one of a lower rank, could be used to explain the data.

We encounter more serious problems when $k \geqslant 1$. In these cases the space of global maxima of the likelihood can be disconnected with up to 8 regions of solutions. We have argued above that we could expect to see such configurations regularly for well-formulated models. A degenerate case, when the possible explanations consist of 8 disconnected points is shown in Fig. 5. This tells us that even if we know the generating mechanism for any exhaustive data set on $[S_4|S_0]$, there will be equally well fitting yet quite different sets of conditional probabilities corresponding to completely different explanations of the data. Such unidentifiability can cause havoc with numerical estimation techniques unless handled with care, for example see [6].

Note that monotonicity constraints used to solve the aliasing problem in Section 4 no longer resolve the ambiguities and multiple maxima in these models.

Finally, when T_v is empty, no hidden variable model corresponds to the maximum likelihood reduced rank model. If, in fact, there is not even a triangle which is close to sandwiching the relevant figures this is strong evidence that the hidden variable model with three states fits poorly. We conjecture that a lower bound for the likelihood ratio of both the hidden variable model and the unconstrained model is now possible, based on the fit of the reduced rank model.

To summarise, even with such a simple factor as the second from Fig. 2, the geometry surrounding the estimation of probabilities is rich and interesting, although the number of hidden states is just 3. An understanding of this geometry can be crucial to estimation since it allows us to find *all* of the well supported models and assess the unidentifiability issues.

6. Conclusions

Algebraic geometry is at the heart of the statistical analysis of discrete Bayesian networks. Gröbner bases are an invaluable tool for highlighting some of the implicit geometrical structure of these statistical models and the shapes of regions of unidentifiability and feasibility, as illustrated in the examples above. However, they do not directly tell the whole story and for a full analysis the bases needs to be used in conjunction with an analysis of the constraints imposed by the convexity properties of probability space. We believe this type of analysis will become generic

to not only to the analysis of discrete DAG models but any statistical models where conditional independence or other algebraic constraints on probabilities play a central role.

Appendix. Derivation of the maximum likelihood rank model estimation algorithm

In the general n–k–m case the relevant Lagrangian is

$$f(\Pi, \Lambda, \boldsymbol{\mu}) = \sum_{i,j} n_{i,j} \log \frac{\bar{\pi}_{i,j}}{\pi_{i,j}} - \sum_{k,l} \lambda_{k,l}\, det(M_{k,l}) - \sum_{i} \mu_i(\pi_{ij} - 1),$$

where $\bar{\Pi}$ is the observed proportion matrix and M_{kl} is the (k,l)th submatrix of Π, that is the 4×4 submatrix with the (k,l)th element of Π as its upper left entry:

$$\frac{\partial f}{\partial \pi_{ij}} = -\frac{n_i j}{\pi_i j} - \left[\sum_{(k,l) \in I(i,j)} \lambda_{k,l} (Adj(M_{k,l})_{j,i}) \right] - \mu_i, \tag{A.1}$$

where $I(i,j) = \{(k,l)\colon \pi_{ij} \in M_{k,l}\}$. By summing (A.1) over j, the adjoint term cancels as this is the expansion by row of a zero determinant. Hence

$$\sum_{j} (-n_{ij} - \hat{\pi}_{ij}\hat{\mu}_i) = 0.$$

But $\sum_j \hat{\pi}_{ij} = 1$ and so

$$\hat{\mu}_i = -\sum_{j} n_i j.$$

Rearranging (A.1), setting to zero and substituting now gives

$$\hat{\pi}_{ij} = \frac{n_{ij}}{\sum_j n_{ij} + \sum_{(k,l) \in I(i,j)} \hat{\lambda}_{k,l} (Adj(M_{kl}))_{ji}}.$$

This expression is intractable however it lends itself naturally to the following iterative scheme which, we conjecture, always converges to the global maximum of the rank model likelihood.

Begin with an initial estimate $\Pi^{(0)}$.

Set

$$(\Pi(\mathbf{C}))_{ij} = \frac{n_{ij}}{\sum_j n_{ij} + \sum_{(k,l) \in I(i,j)} c_{k,l} (Adj(M_{kl}))_{ji}},$$

where $C = (c_{kl})$ is a $(n-k) \times (m-k)$ matrix of dummy variables.

Now set $M_{kl}(\mathbf{C}) = (k,l)$th submatrix of $\Pi(\mathbf{C})$.

Form the system of polynomial equations in the $\mathbf{C}$ variables using the determinant constraint $det(M_{kl}(\mathbf{C})) = 0$ for each (k,l).

Solve the system using Gröbner basis techniques for a solution $\mathbf{C}^*$.

Set $\Pi^{(1)} = \Pi(\mathbf{C}^*)$. Iterate.

References

[1] M.-H. Chen, Q.-M. Shoa, J.G. Ibrahim, Monte Carlo Methods in Bayesian Computation, Springer, New York, 2000.
[2] R.G. Cowell, Mixture reduction via predictive scores, Statist. Comput. 8 (1998) 97–103.
[3] R.G. Cowell, A.P. Dawid, S.L. Lauritzen, D.J. Spiegelhalter, Probabilistic Networks and Expert Systems, Springer, New York, 1999.
[4] R.G. Cowell, P. Sebastiani, A comparison of sequential learning methods for incomplete data, Bayesian Statist. 5 (1996) 533–542.
[5] D.R. Cox, N. Wermuth, Multivariate dependencies, Chapman & Hall, London, 1996.
[6] J. Croft, Estimation in rank models for two-way contingency tables, Research Report 374, University of Warwick, 2000.
[7] J. de Leeuw, P.G.M. van der Heijden, P. Verboon, A latent time-budget model, Statist. Neerlandica 44 (1991) 1–21.
[8] P. Diaconis, B. Sturmfels, Algebraic algorithms for sampling from conditional distributions, Ann. Statist. 26 (1998) 363–397.
[9] I.H. Dinwoodie, The Diaconis–Sturmfels algorithm and rules of succession, Bernoulli 4 (1998) 401–410.
[10] K. Dohmen, Improved inclusion/exlusion identities and inequalities based on a particular class of abstract tubes, Electron. J. Probab. 4 (1999) 12pp.
[11] J.M. Evans, Z. Gilula, I. Guttman, Latent class analysis of two-way contingency tables by Bayesian methods, Biometrika 76 (1989) 557–563.
[12] D. Geiger, D. Heckerman, C. Meek, Asymptotic model selection for directed networks with hidden variables, Proceedings of the 12th Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann, San Mateo, CA, 1996, pp. 283–290.
[13] B. Giglio, D.Q. Naiman, H.P. Wynn, Gröbner bases, abstract tubes and inclusion-exclusion reliability bounds, SCU Research Report 25, Department of Statistics, University of Warwick, 2000.
[14] B. Giglio, E. Riccomagno, H.P. Wynn, Gröbner basis strategies in regression, J. Appl. Statist. 27 (2000) 923–938.
[15] L.A. Goodman, The analysis of systems of qualitative variables when some of the variables are unobserved. A modified latent structure approach, Amer. J. Sociol. 79 (1974) 1179–1259.
[16] L.A. Goodman, Explanatory latent structure analysis using both identifiable and unidentifiable models, Biometrika 61 (1974) 215–231.
[17] S.J. Haberman, Z. Gilula, Canonical analysis of contingency tables by maximum likelihood, J. Amer. Statist. Assoc. 81 (1986) 780–788.
[18] D. Geiger, D. Heckerman, H. King, C. Meek, Stratified exponential families: graphical models and model selection, Technical Report MSR-TR-98-31, Microsoft Research centre, WA, USA, 1998.
[19] M.L. Jordan, Learning in Graphical Models, MIT Press, Cambridge, MA, 1998.
[20] M. Kreuzer, L. Robbiano, Computational Commutative Algebra 1, Springer, Berlin, Heidelberg, 2000.
[21] S.L. Lauritzen, Graphical Models, Oxford University Press, Oxford, 1996.
[22] D. Mond, J.Q. Smith, D. van Straten, Sandwiched triangles, stochastic factorisation and the topology of the space of explanations, JRSS ser. A, 2003, submitted.
[23] K.P. Murphy, A variational approximation for Bayesian networks with discrete and continuous latent variables, Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann, San Mateo, CA, 1999, pp. 457–466.
[24] D. Naiman, H.P. Wynn, Abstract tubes for simplex and orthant arrangements with applications to reliability bounds, SCU Research Report 24, Department of Statistics, University of Warwick, 2000.
[25] G. Pistone, E. Riccomagno, H.P. Wynn, Gröbner bases and factorisation in discrete probability and Bayes, Computing and Statistics: Special Issue for the Workshop on Symbolic Computation, 1997.
[26] G. Pistone, E. Riccomagno, H.P. Wynn, Algebraic Statistics, Chapman & Hall, London, 2001.
[27] G. Pistone, H.P. Wynn, Finitely generated cumulants, Statist. Sinica 9 (1999) 1029–1052.

[28] M. Ramoni, P. Sebastiani, Learning Bayesian networks from incomplete databases, Proceedings of the 13th Conference on Uncertainty in Artificial Intelligence Morgan Kaufmann, San Mateo, CA, 1997, pp. 401–408.
[29] T. Richardson, H. Bailer, M. Banjeree, Tractable structure search in the presence of latent variables, Proc. Artif. Intell. Statist. 1999 (2000) 142–151.
[30] G. Schwartz, Estimating the dimension of a model, Ann. Statist. 6 (1978) 461–464.
[31] R. Settimi, J.Q. Smith, Geometry, moments and conditional independence trees with hidden variables, Ann. Statist. 28 (2000) 1179–1205.
[32] R. Settimi, J.Q. Smith, On the geometry and model selection of Bayesian directed graphs with isolated hidden nodes, Statistica Sinica, 2003, to appear.
[33] D.J. Spiegelhalter, R.G. Cowell, Learning in probabilistic expert systems, Bayesian Statist. 4 (1992) 447–466.
[34] D.J. Spiegelhalter, A.P. Dawid, S.L. Lauritzen, R.G. Cowell, Bayesian analysis in expert systems, Statist. Sci. 8 (1993) 219–282.
[35] D.J. Spiegelhalter, S.L. Lauritzen, Sequential updating of conditional probabilities on directed graph structures, Networks 20 (1990) 579–605.
[36] P. Spirtes, C. Glymour, R. Scheines, Causation, Prediction and Search, in: Lectures Notes in Statistics, Vol. 81, Springer, New York, 1993.
[37] M. Stevens, Dealing with label-switching in mixture models, J. Roy. Statist. Soc. Ser. B 62 (2000) 795–809.
[38] L.A. van der Ark, P.G.M. van der Heijden, Graphical display of latent budget analysis and latent class analysis, with special reference to correspondence analysis, Visualisation of Categorical Data, Academic Press, San Diego, 1998, pp. 489–509.
[39] J. Whittaker, Graphical Models in Applied Statistics, Oxford University Press, Oxford, 1990.

Further reading

D.R. Cox, J. Little, D. O'Shea, Ideals, Varieties, and Algorithms, Springer, New York, 1997.
D.R. Cox, J. Little, D. O'Shea, Using Algebra, Springer, New York, 1998.